ABSTRACT
              This paper presents a new multivariate index for use
in educational planning. This new index, is called a propinquity
index. An individual's propinquity index with respect to a given
occupational group is his geometric distance in n-dimensional space
from the group centroid, where each dimension is the standard score
on one of the original variables weighted by a value representing, at
least approximately, the relevance of the corresponding variable in
identifying group members. The propinquity index may be used in two
ways in helping the individual develop his educational and vocational
plans. Converted to a propinquile, it can constitute one of many
separate items of information used in arriving at important
decisions. In many circumstances the propinquity index may function
better as one of the predictors in a multiple regression equation to
predict a dichotomous group membership criterion. (Author/KJ)

A New Multivariate Index for Use in Educational Planning*

Marion F. Shaycoft
American Institutes for Research
Palo Alto, California

The late Phillip Rulon once wrote a delightful little fable about a Hexagon who didn't know whether it should become a Circle or a Square - because it had corners, like a Square, but was round like a Circle. The Hexagon was given conflicting advice by a Relentless Psychologist who told it it would be a superior Circle, and an Inveterate Statistician who told it it was really a Square, not a Circle at all. (Rulon wrote his article before the term "Square" had the connotation it has today.) The Inveterate Statistician used Discriminant Analysis to come to his conclusion, while the Relentless Psychologist used Stanines for the purpose. Rulon ended his fable with the following "moral":

> "The Multiple Correlation Technique applied successively
> to different groups yields information not given by the
> Discriminant Function applied to all the groups,
>
> and vice versa."

The moral is perfectly sound, but unfortunately the resulting conflicting advice left the poor little Hexagon in a state of perplexity, not knowing whether to get two of its corners rubbed off and become a Square or get all of its corners rubbed off and become a Circle.

What might have been added to the moral is that the multiple correlation approach and the multiple discriminant function approach, even in combination, don't yield all the relevant information available.

The term "multiple correlation approach", in this context, refers to

---

the correlation of a set of predictors against a continuous criterion that distinguishes among members of the group (in terms of quality of performance or whatever other characteristic seems relevant). The term "multiple discriminant function approach" as used here refers to the situation where the total sample has been divided into several groups and the sole criterion is membership in a group. There is no continuous criterion available to distinguish among members of the same group. The discriminant functions are a set of linear composites of the original predictor variables, and considered in combination these discriminant functions locate the group centroids in a space which includes all dimensions in which they differ, and maximizes the between-groups variance relative to the within-groups variance, in that space.

Actually in addition to these two approaches there is a third commonly used one, which is really a cross between the first two. This is the approach where even though there are more than two groups, instead of performing one multiple discriminant analysis that takes all groups into account simultaneously, a separate discriminant analysis is performed for each group, to discriminate between that group and all other groups combined. It is well known, of course, that this results in a single discriminant function for each group, and that this discriminant function is identical (except for its standard deviation) with the multiple regression equation against the dichotomous criterion of group membership. In other words it is equivalent to the results of a conventional multiple correlation approach where point biserial correlations are used for the predictor-criterion correlation coefficients. I shall therefore refer to this as the "multiple point biserial R" approach, by way of shorthand.

In addition to these three techniques (multiple correlation, multiple discriminant analysis, and multiple point biserial R), I am proposing a fourth technique, called "propinquity analysis", which superficially has some of the features of each of the other approaches but is actually quite different from any of them.

Before I go into the details of this procedure, let me tell you
what Rulon's hexagon should become according to the results of propinquity
analysis. At this meeting of psychologists I hate to disagree with the
"Relentless Psychologist", but the fact is that the Hexagon resembles
a Square much more closely than it resembles a Circle. Its propinquile
for being a Square is about 85, while its propinquile for being a
Circle is less than 10.

A propinquile, I should explain, is the percentile corresponding
to the propinquity index.

THE INADEQUACIES OF CONVENTIONAL METHODS

Perhaps the simplest way to describe what the propinquity index is and
does is to start with a discussion of the problems we have to bypass - the
problems and disadvantages inherent in any guidance system that depends
solely or heavily on discriminant functions. I want to stress that what
we are talking about in this connection is the guidance problem, not the
personnel classification problem. Most of the difficulties and disadvan-
tages in connection with multiple discriminant functions are of little or
no relevance in connection with personnel classification, where there are
specified numbers of vacancies in various occupational categories, with
the total number of vacancies in all categories combined exactly matching
the number of people to be assigned to the vacancies. The multiple
discriminant function approach of course has many other appropriate uses,
besides personnel classification. But here are some of the disadvantages
of using it as the basis for guidance:

1. If more than one discriminant function is involved, which is

almost certain to be the case if there are at least two variables
and at least three groups, generally each of these composite
variables except the first is bipolar, having some positive and
some negative correlations with the original variables. Moreover
the set of "regression weights" applied to the original variables
to compute any discriminant function beyond the first generally
includes both positive and negative values of substantial magnitude.
All of this is likely to make discriminant functions beyond the
first very obscure in meaning - if not downright uninterpretable -
at least in the realm of cognitive variables, where the concept
of a truly bipolar factor seems inherently somewhat contradictory
and therefore not very useful. Discriminant functions are
admittedly artifacts, and there is little reason for using
artifacts that are not basically helpful, merely because they
happen to have some interesting statistical properties. The
fact that discriminant functions maximize a certain kind of
variance ratio when computed from a specified set of variables,
with respect to a specified set of groups, doesn't automatically
endow all discriminant functions with intrinsic importance or
make them good tools for guidance.

2. But if bipolar discriminant functions are hard to interpret
and give names to, how much harder it must be to try to
explain them to someone else - someone vitally concerned
with what they really mean. For instance it seems to me
that it would be awfully difficult to explain to a boy who
has his heart set on becoming a lawyer that he really shouldn't
because his score is too low on a composite composed of (to
take a strictly hypothetical example) two parts "Arithmetic
Reasoning" plus one part "Memory for Words" minus one part
"Spelling" plus two parts "English Usage" minus three parts
"Visualization in Three Dimensions" ..... etc.

3. And what helpful suggestions can one give the would-be lawyer? That he should acquire the habit of misspelling, since the lower his Spelling score is the less will be subtracted? Or merely that the next time he takes such a battery he should deliberately give incorrect answers on those tests he knows to be negatively weighted on the discriminant function on which he wants to improve his score?

4. To get an idea of some even more serious disadvantages, look at Figure 1, on page 19 (in the Appendix). In this diagram the two axes, X_1 and X_2, represent scores on two uncorrelated variables with equal standard deviations. The large circle represents the total group, T. The center of the circle is the centroid of the group and the radius equals 2σ so that almost everyone in the group is within the circle. Likewise the two smaller solid circles represent subgroups, A and B. The total group, T, consist of three subgroups -- A, B, and an amorphous group containing everybody else. Since the centroids of A, B, and T are collinear, the centroid of the amorphous residual group, T-A-B, is necessarily also collinear with the A and B centroids; therefore in a discriminant analysis of Groups, A, B, and the residual group there would be only one discriminant function, which would be a rescaling of Variable X_1. The Variable X_2 data would be lost, although the fact that there is a marked restriction of Group A's range with respect to Variable X_2 suggests that that variable has some importance in determining a person's suitability for Group A. (To a lesser extent the same thing is true of Group B.) Individual I_1, who seems seriously deficient on whatever quality is measured by X_2, would be diagnosed as fitting into Group A optimally, if the discriminant analysis of Groups A, B, and the residual group were the sole basis of the decision. Individual I_2, who is apparently very much overqualified on X_2, would likewise be assigned to Group A. Thus discriminant analysis may produce a set of composites which omits one or more important dimensions.

5. Now let's look at the two dotted circles, C and D. Suppose
   that these are two additional groups, separated out from the
   residual group, and that a five-group multiple discriminant
   analysis is now done, on Groups A, B, C, D, and what remains
   of T. Because Groups C and D have entered the picture a
   second discriminant function, which is a rescaling of X_2, appears.
   It then becomes evident that cases I_1 and I_2 would be misfits
   in Group A despite their fine fit on the first discriminant
   function, because they are so far off on the second. But it
   seems unreasonable that an evaluation of whether one is well
   suited for a particular category should depend on whether certain
   other groups - ones that are definitely inappropriate - have been
   separated out.

The propinquity index and its close relative the propinquile were
designed to be free of all these problems inherent in the multiple dis-
criminant analysis approach. But there are also problems in connection
with the multiple correlation and multiple point biserial R, and the
propinquity statistics were designed to avoid those problems, too.
Inherent in the multiple point biserial R approach (and also in the
multiple correlation approach) is the assumption that the more one
possesses of an ability or other trait that has a positive beta weight,
the more suited one is for membership in the occupational group (or
whatever other kind of group is involved), and that the more one possesses
of an ability or trait that has a negative beta weight, the less suited
one is for group membership.

Consider what this means in terms of an occupation which requires
about average intelligence. If the incumbents in this group happen to
have an average IQ of 98 and if the range of IQ's for this group happens
to be between 90 and 107 (which is quite a restricted range), the beta

weight is likely to be a small <u>negative</u> number, and someone with an IQ of 70 will appear more qualified than someone with an IQ of 100.

Now suppose that instead of a mean IQ of 98 the incumbents happen to have a mean IQ of 102, but the same restricted range. The beta weight for this situation is likely to be a small <u>positive</u> number, and someone with an IQ of 150 will appear to be better qualified than someone with an IQ of 101, when actually the person with an IQ of 150 would be grossly overqualified. And in view of the resultant waste of talent, overqualification for an occupation is about as undesirable as gross underqualification - at least it is in the square world of the non-hippy non-dropout-from society.

Now let's look at the third possibility - that the mean IQ for the group is very close to 100, still with a very restricted range, and that the beta weight happens to be 0. In this situation neither the person with an IQ of 150 nor the one with an IQ of 70 would be disqualified, although the restriction of range suggests that both should be.

The propinquity statistics take care of all these problems.

THE NATURE AND CHARACTERISTICS OF PROPINQUITY INDEXES

Briefly, an individual's propinquity index with respect to a given occupational group is his geometric distance* in n-dimensional space from the group centroid, where each dimension is the standard score on one of the original variables (standardized in terms of the group in question) weighted by a value representing, at least approximately, the relevance of the corresponding variable in identifying group members. A minus sign is attached to the distance, so that a propinquity index of 0, indicating that the individual's scores on relevant variables coincide with the group centroid, is the maximum value of the index.

---

*This represents a very slight change from the original definition, which was used in the oral presentation at APA. According to the original definition, the propinquity index was the <u>square</u> of the distance. The reason for the change was to improve some statistical characteristics of the index.

The greater the distance, the larger the absolute value of the index. The
purpose of the minus sign, therefore, is to orient the propinquity index
properly, so that when it is used in a correlation matrix there won't
be inconvenient negative correlations. The higher the algebraic value
of the index, the <u>closer</u> the individual is to the centroid. (Hence the
term "propinquity index".)

Formula 12 (or 13) in the appendix* represents the squared distance in
regard to a single variable. Formula 14 gives the propinquity index, δ, with
w as the weight representing the relevance of a particular variable.

Formulas 22-28 are seven different formulas giving different results,
that might be used for determining the w's. All these formulas have the
desired characteristic of giving a weight of 0 for irrelevant variables
and a positive weight for relevant variables. There are no negative
weights. The indicator of relevance for a variable is a function of the
ratio of the standard deviation within the group to the standard deviation
of the total sample (all groups combined).

Empirical exploration is necessary in order to decide which of the
seven formulas for w_{ij} is the one that should be used. It might occur
to some of you that instead of using any of them we should just use
multiple regression weights (against the dichotomous criterion of group
membership). This is a possibility, but there are several factors
militating against it. In the first place there is the bouncing beta
phenomenon, which is likely to be a problem when the number of predictors
is large -- even if there is a very large number of cases. And the fact
that the predictors are not merely linear functions of the variable,
but involve squared terms, might make the betas bounce even more. Another
disadvantage of regression weights lies in the very practical consideration
that even with a high-powered computer, the computation of the betas for
these squared terms is an extraordinarily complex, time-consuming, and
expensive operation when the number of cases, number of variables, and

---

*All formulas in this paper are in the Appendix, which also contains a
 section defining all the notation used.

number of groups are all comparatively large (as is true in the case of
the Project TALENT data, to which this technique is being applied).

In view of the practical and theoretical objections to the use of multiple
regression weights in this context, the best solution seems to be to see which
of the seven systems of weights (Formulas 22-28) gives the weighted composites
that have the highest point biserial correlation with the dichotomous group
membership criterion.

All seven of the possible formulas for weights use some function of
$\sigma_{z_{ij}}$ as the indicator of relevance of the variable. Since $\sigma_{z_{ij}}$ is the
ratio of the standard deviation within the group to the total standard
deviation (Formula 15) there is an assumption that a relevant variable
will have a somewhat restricted range, and the more relevant the variable,
the more restricted the range. This assumption may not be sound when
some of the variables have skewed distributions. For such variables
the ratio of group σ to total σ may be erratic, resulting in peculiar
weights (w). Therefore it is desirable that all variables used in a
propinquity analyses have normal distributions. If there is no reason
to believe that the distributions are at least approximately normal they
should be normalized.

Because the variables used in computing the propinquity index are
raw scores rather than principal components, discriminant functions, or
some other kind of uncorrelated composites, propinquity indexes may
lack some of the mathematical precision and invariance of statistics
computed in a geometric space where orthogonal axes correspond to un-
correlated variables. But this seems a small price to pay for the twin
advantages of interpretability and ready explainability. Of course
principal components or discriminant functions could be used as the
initial scores, but it isn't recommended.

PROPINQUILES

In the univariate case the distribution of δ resembles the left half of a normal distribution, assuming the original variable is normally distributed. But when the number of variables entering into the propinquity index is greater than one, neither the distribution of δ nor the distribution of δ^2 is normal, nor are these distributions any portion of a normal distribution. Furthermore the basic shape of the distribution varies with the number of variables involved. Propinquity indexes based on three variables have an entirely different distribution from those based on two, and so forth.

And even when propinquity indexes for a set of groups are all based on the same battery they still may not be directly comparable, because though all indexes technically are based on the same number of variables, some variables may have zero weights for certain groups so that the number of dimensions is in effect different for different groups. Since propinquity indexes for different categories therefore are not generally directly comparable, they have to be converted to some uniform scale in order to be compared - and percentiles serve this purpose effectively and directly. These percentiles are the propinquiles.

THE ROLE OF PROPINQUITY ANALYSIS

Propinquity analysis is not in any sense of the word a replacement for the multiple correlation approach, nor for the multiple point biserial R approach. Rather, propinquity analysis should serve as an adjunct to both these procedures. In situations where the group membership dichotomous criterion is available, not only can the two procedures, propinquity analysis and multiple point biserial R, be used independently, but also the multiple point biserial R approach can be modified to

incorporate the propinquity index as an additional predictor variable.
This couldn't be done, of course, if the propinquity index were merely
some sort of linear function of the n initial predictors, but it isn't,
since squared terms are also involved. The propinquity index thus is
not linearly dependent on the n variables which enter into it.

A few special considerations should be pointed out in regard to
the use of the propinquity index as an additional predictor:

1. In the first place, the weights for the U components of the
   propinquity index should probably be determined on the basis
   of a different sample from the one used for the correlation
   matrix, in order to avoid capitalizing on chance in the point
   biserial correlation of the propinquity index with the dichotomous
   criterion, and to prevent this correlation from being spuriously
   high. The same thing applies if biserial correlations rather
   than point biserials are used, although it is not advisable to
   do this unless all the predictive variables, including the
   propinquity index, either have been normalized or are known to
   be normally distributed. Since two separate (but parallel) samples
   should be used on the combination procedure, a third parallel
   sample should also have been set aside initially for cross-
   validation.*

2. To use this combination procedure, a very large number of cases
   has to be available, not only because the initial sample has
   to be split into three separate samples in order to carry out
   the complete procedure properly, including cross-validation, but

---

*The point biserial correlations of the dichotomous criterion with the n
individual variables and with the propinquity index are of course part
of the correlation matrix and should therefore be based on the same
sample.

also because of the bouncing beta problem and because the betas
may bounce especially erratically when a variable involving
squares is involved, such as the propinquity index, the sample
used for the multiple correlation analysis should probably be
substantially larger than either of the other two, to control
as well as possible this instability.

3. The propinquity index should be retained in the final regression
equation only for those dichotomous criteria where it seems to
add significantly to the cross-validated multiple correlation.
This is probably _most_ likely to occur in situations where both
overqualification and underqualification are serious deterrents
to group membership - rather than the somewhat more usual
situation where underqualification is an overwhelmingly more
potent deterrent than overqualification is. In the vocational
planning or vocational guidance situation this is more likely
to happen in the case of middle-level occupations, and perhaps
even, to a certain extent, in the case of lower-level occupations,
than in the case of occupations that demand very high levels of
ability. In the case of these more demanding occupations, the
higher one's qualifications are the better.

CONCLUSIONS

In summary, the propinquity index may be used in two ways in helping
the individual develop his educational and vocational plans. Converted
to a propinquile, it can constitute one of many separate items of
information used in arriving at important decisions. Or in many circum-
stances the propinquity index may function better as one of the predictors
in a multiple regression equation to predict a dichotomous group member-
ship criterion.

At the risk of oversimplifying horribly - and I <u>know</u> that I am -
I would like to close by suggesting that perhaps the principal statistical
approaches that should be used in three fields of application - selection,
classification, and guidance - are as follows:

For personnel classification:	Multiple discriminant function analysis.
For personnel selection:	Multiple correlation approach.
For guidance:	Propinquiles, and the multiple point biserial R approach, modified to include the propinquity index as one of the predictors.

REFERENCES

Rulon, P. J. "The Stanine and the Separile: A Fable", <u>Personnel Psychology</u>,
<u>4</u>, 1951, 99-114.

Rulon, P. J. "Distinctions between Discriminant and Regression Analyses
and a Geometric Interpretation of the Discriminant Function", in "The
Multiple Discriminant Function - A Symposium". <u>Harvard Educational
Review</u>, vol. XXI, No. 2, Spring 1951, 80-90.

Tiedeman, D. V. "The Utility of the Discriminant Function in Psychological
and Guidance Investigations", in "The Multiple Discriminant Function -
A Symposium", <u>Harvard Educational Review</u>, vol. XXI, No. 2, Spring 1951, 71-80.

APPENDIX

I. NOTATION

n = no. of variables (≈ no. of dimensions)

g = no. of groups

N_j = no. of cases in group j

N = total no. of cases

$$N = \sum_{j=1}^{g} N_j \tag{1}$$

X_{ik} = raw score of individual k on variable i

$$i = 1,2,3,\ldots n$$
$$k = 1,2,3,\ldots N_j$$

$\overline{X}_i$ = mean of variable X_i for total group

$$\overline{X}_i = \frac{\sum_{k=1}^{N} X_{ik}}{N} \tag{2}$$

s_{x_i} = sample standard deviation of variable X_i for total group

$$s_{x_i} = \sqrt{\frac{\sum_{k=1}^{N} (X_{ik} - \overline{X}_i)^2}{N}} \tag{3}$$

σ_{x_i} = corresponding estimate of population standard deviation

$$\sigma_{x_i} = s_{x_i} \sqrt{\frac{N}{N-1}} \tag{4}$$

z_{ik} = standard score of individual k on variable i

$$z_{ik} = \frac{X_{ik} - \overline{X}_i}{\sigma_{x_i}} \tag{5}$$

$$\overline{z}_i = 0 \tag{6}$$

Note that σ_x not s_x is in the denominator of z_{ik} in this algebraic development.

$$\therefore \quad s_{z_i} = \sqrt{\frac{N-1}{N}} \tag{7}$$

$$\sigma_{z_i} = 1 \tag{8}$$

$\overline{X}_{ij}$ = means of variable X_i for group j

$$\overline{X}_{ij} = \frac{\sum\limits_{k=1}^{N_j} X_{ik}}{N_j} \qquad (9)$$

$$i = 1,2,3,\ldots n$$
$$j = 1,2,3,\ldots g$$

$s_{x_{ij}}$ = sample standard deviation of variable i for group j

$$s_{x_{ij}} = \sqrt{\frac{\sum\limits_{k=1}^{N_j} (X_{ik} - \overline{X}_{ij})^2}{N_j}} \qquad (10)$$

$\sigma_{x_{ij}}$ = estimate of population standard deviation of variable i for group j

$$\sigma_{x_{ij}} = s_{x_{ij}} \sqrt{\frac{N_j}{N_j - 1}} \qquad (11)$$

U_{ijk} = propinquity component indicating how much like group j individual k is, with respect to variable i.

w_{ij} = weight representing the relevance of component U_i as an indicator of membership of individual in group j.

δ_{jk} = propinquity index for individual k indicating how much like group j he is.

P_{jk} = propinquile for individual k corresponding to his propinquity index for group j. The propinquile is a percentile based on the individuals in group j. A propinquile of 100 corresponds to a propinquity index of 0.

II. FORMULAS

$$U_{ijk} = \left(\frac{z_{ik} - \overline{z}_{ij}}{\sigma_{z_{ij}}} \right)^2 \qquad (12)$$

$$= \left(\frac{X_{ik} - \overline{X}_{ij}}{\sigma_{x_{ij}}} \right)^2 . \qquad (13)$$

$$\delta_{jk} = -\sqrt{\sum_{i=1}^{n} w_{ij} U_{ijk}} \qquad (14)$$

$$\sigma_{z_{ij}} = \frac{\sigma_{x_{ij}}}{\sigma_{x_i}} \qquad (15)$$

$$= \frac{\sigma_{x_{ij}}}{s_{x_i}} \sqrt{\frac{N-1}{N}} \qquad (16)$$

Formulas for intermediate values in computing w_{ij}

$$w_1' = \log_{10} \frac{1}{\sigma_{z_{ij}}} \tag{17}$$

$$w_2' = \frac{1}{\sigma_{z_{ij}}} - 1 \tag{18}$$

$$w_3' = 1 - \sigma_{z_{ij}} \tag{19}$$

$$w_4' = \frac{1}{\sigma^2_{z_{ij}}} - 1 \tag{20}$$

$$w_5' = 1 - \sigma^2_{z_{ij}} \tag{21}$$

Formulas for w_{ij}

$$w_1 = w_1' \quad \text{with negative values changed to 0} \tag{22}$$

$$w_2 = w_2' \quad " \quad " \quad " \quad " \quad " \ " \tag{23}$$

$$w_3 = w_3' \quad " \quad " \quad " \quad " \quad " \ " \tag{24}$$
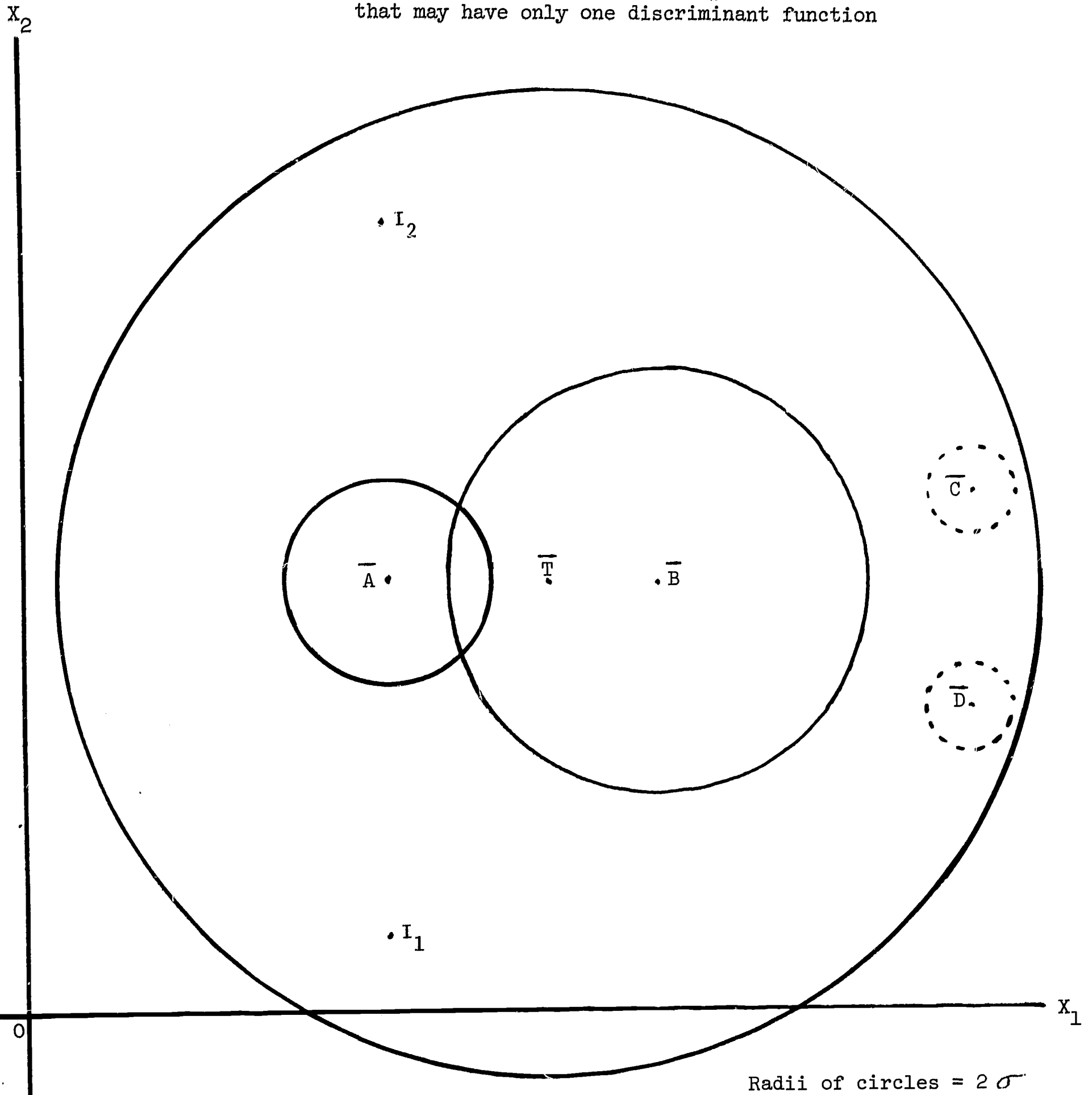
$$w_4 = w_4' \quad " \quad " \quad " \quad " \quad " \ " \tag{25}$$

$$w_5 = w_5' \quad " \quad " \quad " \quad " \quad " \ " \tag{26}$$

$$w_6 = w_2^2 \tag{27}$$

$$w_7 = w_3^2 \tag{28}$$

FIGURE 1. Illustrating data with two important dimensions
that may have only one discriminant function

Groups A, B, and T-A-B have only one discriminant function.
When Groups C and D are added, a second discriminant function appears.

I_1 and I_2 are individuals classified on the basis of discriminant
analysis as perfect fits in Group A (unless Groups C and D have
been included in the analysis), but propinquity analysis definitely
excludes them from Group A.